European Language Resource Coordination
Connecting Europe Facility

European Commission

The new Anonymisation Specification at a glance

Tom Vanallemeersch (CrossLang)

CONTEXT

Organisations (public, commercial, research) want to **archive and share data**

➢ Also **multilingual** distribution of data

Examples:

- EC's Digital Service Infrastructures (collaboration with Member States)
- Country Profiles in ELRC White Paper

How to avoid violation of GDPR ?

➢ Removal of confidential data, e.g. names, account numbers

➢ Deidentification: ensure data cannot be associated with any individual, organisation

European Language Resource Coordination
Connecting Europe Facility

European Commission

Digital Service Infrastructures

Needs:

- e-Justice: publication of case law
- ODR (Online Dispute Resolution): consumer complaints
- Europeana: user logs
- Safer Internet: reports on abuse

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC White Paper

- France: development of own NMT solution by some stakeholders
- Italy: upload of potentially confidential or personal data to public MT interfaces
- Norway: lack of awareness from external executives dealing with translation memories
- Sweden: in-house translation services

PURPOSE OF SPECIFICATION

Create technical procedures and best practices for automated anonymisation

- Monolingual setting
- Multilingual setting (MT, translation memories)
- Focus on unstructured data (running text)
- Standardisation, interoperability

Collaborate with other projects

- MAPA (Multilingual Anonymisation toolkit for Public Administrations, CEF)
- ELG (European Language Grid, H2020): NER, privacy preservation

ORGANISATION

1. Consultation round with stakeholders
   - ➢ Understand their practices and needs
   - ➢ Apply bottom-up approach
2. Set up draft specification
3. Feedback from stakeholders
4. Set up final specification
   - ➢ Technical procedures, best practices
   - ➢ Multilingual extension of annotation scheme
   - ➢ Proof-of-concept pipeline (potential workflows)

CONSULTATION ROUND

- Consortium of MAPA
- eTranslation development team at DG Translation: MT, NER
- Domain experts
  - University of Bologna: legislative documents
  - Vicomtech: health data
  - University College London: police reports
- Company SDL: anonymisation tools for translation projects, memories
- Members of ELG consortium and Community
- Language Resource Board of ELRC: *the present meeting*

FINDINGS: USABILITY OF ANONYMISED DATA

- Data sensitivity differs according to domain
  - ➢ Legal domain, police reports, medical data, consumer complaints, …
- There is a trade-off between extent of anonymisation and need for information
  - ➢ Aim for readability or for downstream task (e.g. MT, creation of statistics, …) ?
  - ➢ Example: replace proper names consistently for readability

FINDINGS: USER ORIENTATION

- Toolkit developers should be transparent about risks to users
- Users need control over the anonymisation process
  - ➢ Select part of documents to anonymise (possibly using machine learning)
  - ➢ (De)select (categories of) named entities to be annotated
  - ➢ (De)select text fragments that have been annotated

FINDINGS: ANONYMISATION PIPELINE

- Named-entity recognition (NER) step
  - ➢Training of deep-learning models (+ pre-trained BERT, cross-lingual transfer)
  - ➢Regular expressions
  - ➢Gazetteers with lists of named entities
- Anonymisation step
  - ➢Mask entities using crosses
  - ➢Replace entities using pseudonym (label, replacing word, encryption string)
- Mapping table for back-mapping (data owner)

NER need not be perfect: make sure anonymisation is undetectable for attackers

FINDINGS: ANNOTATION PROCESS

- Toolkit should be flexible in terms of annotation categories, hierarchy
  - ➢Cfr. XML in MAPA
- Annotation is sped up using bootstrapping and cross-lingual transfer
- Anonymised metadata (document-level, sentence-level) is also interesting to store
- There is a need for adding a translation layer (nondestructive annotation)
  - ➢Inspiration from XLIFF ?
- Anonymising MT training data and input improves MT and addresses privacy concerns
  - ➢Some organisations want to anonymise data themselves before MT is trained/applied

FINDINGS: ANNOTATION PROCESS

Annotation in the INCEpTION tool used by MAPA:

FINDINGS: EVALUATION

- NER is evaluated using a gold standard
- The evaluation of the anonymisation step is domain-dependent
  - *Potential clues in context even when named entities are correctly annotated*
  - Specific test: motivated intruder test
  - Need for domain expertise
  - Focus on false positives rather than false negatives
- *Anonymisation in the legal sense ≠ anonymisation in the technical sense !*

European Language Resource Coordination
Connecting Europe Facility

European Commission

DISCUSSION

European Language Resource Coordination
Connecting Europe Facility

European Commission

THANK YOU FOR YOUR ATTENTION!

Website: www.lr-coordination.eu

Twitter: @LR_Coordination

Email: info@lr-coordination.eu